writing of more than two distinct states within each memory cell over an extended lifetime of the memory cells, so that more than one bit may be reliably stored in each cell.

According to one aspect of the present invention, the multiple threshold breakpoint levels are provided by a set of memory cells which serves as master reference cells. The master reference cells are independently and externally programmable, either by the memory manufacturer or the user. This feature provides maximum flexibility, allowing the breakpoint thresholds to be individually set within the threshold window of the device at any time. Also, by virtue of being an identical device as that of the memory cells, the reference cells closely track the same variations due to manufacturing processes, operating conditions and device aging. The independent programmability of each breakpoint threshold level allows optimization and fine-tuning of the threshold window's partitioning, critical in multi-state implementation. Furthermore, it allows post-manufacture configuration for either 2-state or multi-state memory from the same device, depending on user need or device characteristics at the time.

A2 Cont

According to another aspect of the present invention, a set of memory cells within each sector (where a sector is a group of memory cells which are all erased at the same time in a Flash EEprom) are set aside as local reference cells. Each set of reference cells tracks the Flash cells in the same sector closely as they are both cycled through the same number of program/erase cycles. Thus, the aging that occurs in the memory cells of a sector after a large number of erase/reprogram cycles is also reflected in the local reference cells. Each time the sector of flash cells is erased and reprogrammed, the set of individual breakpoint threshold levels are re-programmed to the associated local reference cells. The threshold levels read from the local reference cells then automatically adjust to changing conditions of the memory cells of the same sector. The threshold window's partitioning is thus optimally maintained. This technique is also useful for a memory that employs only a single reference cell that is used to read two state (1 bit) memory cells.

According to another aspect of the present invention, the threshold levels rewritten at each cycle to the local reference cells are obtained from a set of master cells which are not cycled along with the memory cells but rather which retain a charge that has been externally programmed (or reprogrammed). Only a single set of master memory cells is needed for an entire memory integrated circuit.

In one embodiment, the read operation directly uses the threshold levels in the local reference cells previously copied from the master reference cells. In another embodiment, the read operation indirectly uses the threshold levels in the local reference cells even though the reading is done relative to the master reference cells. It does this by first reading the local reference cells relative to the master reference cells. The differences detected are used to offset subsequent regular readings of memory cells relative to the master reference cells so that the biased readings are effectively relative to the local reference cells.

According to another aspect of the present invention, the program and verify operations are performed on a chunk (i.e. several bytes) of addressed cells at a time. Furthermore, the verify operation is performed by circuits on the EEprom chip. This avoids delays in shipping data off chip serially for verification in between each programming step.

A2 concl.

According to another aspect of the present invention, where a programmed state is obtained by repetitive steps of programming and verifying from the "erased" state, a circuit verifies the programmed state after each programming step with the intended state and selectively inhibits further programming of any cells in the chunk that have been verified to have been programmed correctly. This enables efficient parallel programming of a chunk of data in a multi-state implementation.

According to another aspect of the present invention, where a chunk of EEprom cells are addressed to be erased in parallel, an erased state is obtained by repetitive steps of erasing and verifying from the existing state to the "erased" state, a circuit verifies the erased state after each erasing step with the "erased" state and selectively inhibits further erasing of any cells in the chunk that have been verified to have been erased correctly. This prevents over-erasing which is stressful to the device and enables efficient parallel erasing of a group of cells.

According to another aspect of the present invention, after a group of cells have been erased to the "erased" state, the cells are re-programmed to the state adjacent the "erased" state. This ensures that each erased cell starts from a well defined state, and also allows each cell to undergo similar program/erase stress.

According to another aspect of the present invention, the voltage supplied to the control gates of the EEprom cells is variable over a wide range and independent of the voltage supplied to the read circuits. This allows accurate program/erase margining as well as use in testing and diagnostics.--

Page 5, line 28, change the period "." to a semicolon --;--, and add the following:

-Figure 9 is a cross-sectional view of an EEprom device integrated circuit structure that can be used to implement the various aspects of the present invention;

Figure 10 is a view of the structure of Figure 9 taken across section 2-2 thereof;

Figure 11 is an equivalent circuit of a single EEprom cell of the type illustrated in Figures 9 and 10;

Figure 12 shows an addressable array of EEprom cells;

Figure 13 is a block diagram of an EEprom system in which the various aspects of the present invention are implemented;

Figure 14 illustrates the partitioning of the threshold window of an EEprom cell which stores one bit of data;

Figure 15A illustrates the partitioning of the threshold window of an EEprom cell which stores two bits of data;

Figure 15B illustrates the partitioning of the source-drain conduction current threshold window of the EEprom cell of figure 15A;

Figures 16A and 16B are curves that illustrate the changes and characteristics of a typical EEprom after a period of use;

Figure 17A illustrates read and program circuits for a master reference cell and an addressed memory cell according to the present invention;

Figure 17B illustrates multi-state read circuits with reference cells according to the present invention;

Figures 17C(1)-17C(8) illustrate the timing for multi-state read for the circuits of Figure 17B;

Figure 18 illustrates a specific memory organization according to the present invention;

Figure 19 shows an algorithm for programming a set of local reference cells according to the present invention;

Figure 20A shows one embodiment of a read circuit using local reference cells directly;

Figure 20B shows a read algorithm for the embodiment of Figure 20A;

Figure 21A shows an alternative embodiment of a read circuit using local reference cells indirectly;

Figure 21B is a programmable circuit for the biased reading of the master reference cells according to the alternative embodiment;

Figure 21C is a detail circuit diagram for the programmable biasing circuit of Figure 21B;

Figure 21D shows a read algorithm for the embodiment of Figure 21A;

Figure 22 illustrates the read/program data paths for a chunk of cells in parallel;

A3 cont'd.

Figure 23 shows an on chip program/verify algorithm according to the present invention;

Figure 24 is a circuit diagram for the compare circuit according to the present invention;

Figure 25 is a circuit diagram for the program circuit with inhibit according to the present invention; and

Figures 26 and 27 are tables that list typical examples of operating voltages for the EEprom cell of the present invention.--

Page 11, line 26, change "Harari" to --Harari, now patent no. 5,095,344,--.

Page 11, lines 28 and 29, strike "filed on the same day as the present application," and substitute the following therefore: --Serial No. 07/337,579, filed April 13, 1989, now abandoned,--.

Page 22, line 14, insert after "204,175" --now patent no. 5,095,344,--.

Page 22, line 16, change "Techniques." to --Techniques, Serial No. 07/337,579, filed April 13, 1989, now abandoned.--

Page 26, line 3, insert a comma --,-- after "204,175" and insert thereafter --now patent no. 5,095,344,--.

Page 26, line 4, strike all of line 4, and substitute the following therefore: --Harari, Serial No. 07/337,579, filed April 13, 1989, now abandoned,--

Page 32, between lines 20 and 21, insert the following:

A4

--There are many specific Eprom, EEprom semiconductor integrated circuit structures that can be utilized in making a memory array with which the various aspects of the present invention are advantageously implemented.

"Split-Channel" EEprom Cell

A preferred EEprom structure is generally illustrated in the integrated circuit cross-sectional views of Figures 9 and 10. Describing this preferred structure briefly, two memory cells 1011 and 1013 are formed on a lightly p-doped substrate 1015. A heavily n-doped implanted region 1017 between the cells 1011 and 1013 serves as a drain for the cell 1011 and a source for the cell 1013. Similarly, another implanted n-doped region 1019 is the source of the cell 1011 and the drain of an adjacent cell, and similarly for another n-doped region 1021.

A4 cont

Each of the memory cells 1011 and 1013 contains respective conductive floating gates 1023 and 1025, generally made of polysilicon material. Each of these floating gates is surrounded by dielectric material so as to be insulated from each other and any other conductive elements of the structure. A control gate 1027 extends across both of the cells 1011 and 1013 in a manner to be insulated from the floating gates and the substrate itself. As shown in Figure 10, conductive strips 1029 and 1031 are additionally provided to be insulated from each other and other conductive elements of the structure, serving as erase gates. A pair of such erase gates surrounds the floating gate of each memory cell and are separated from it by an erase dielectric layer. The cells are isolated by thick field oxide regions, such as regions 1033, 1035, and 1037, shown in the cross-section of Figure 9, and regions 1039 and 1041 shown in the view of Figure 10.

The memory cell is programmed by transferring electrons from the substrate 1015 to a floating gate, such as the floating gate 1025 of the memory cell 1013. The charge on the floating gate 1025 is increased by electrons traveling across the dielectric from a heavily p-doped region 1043 and onto the floating gate. Charge is removed from the floating gate through the dielectric between it and the erase gates 1029 and 1031. This preferred EEprom structure, and a process for manufacturing it, are described in detail in copending patent application Serial No. 323,779 of Jack H. Yuan and Eliyahou Harari, filed March 15, 1989, which is expressly incorporated herein by reference.

The EEprom structure illustrated in Figures 9 and 10 is a "split-channel" type. Each cell may be viewed as a composite transistor consisting of two transistor T1 and T2 in series as shown in Figure 11. The T1 transistor 1011a is formed along the length L1 of the channel of the cell 1011 of Figure 9. It has a variable threshold voltage V_{T1}. In series with the T1 transistor 1011a is the T2 transistor 1011b that is formed in a portion of the channel L2. It has a fixed threshold voltage V_{T2}

of about 1V. Elements of the equivalent circuit of Figure 11 are labeled with the same reference numbers as used for corresponding parts in Figures 9 and 10, with a prime (') added.

As can best be seen from the equivalent circuit of Figure 11, the level of charge on the T1's floating gate 1023' of an EEprom cell affects the threshold voltage V_{T1} of the T1 transistor 1011a when operated with the control gate 1027'. Thus, a number of memory states may be defined in a cell, corresponding to well defined threshold voltages programmed into the cell by an appropriate amount of charge placed on the floating gate. The programming is performed by applying, over a certain period of time, appropriate voltages to the cell's control gate 1027' as well as drain 1017' and source 1019'.

Addressable Flash EEprom Array

The various aspects of the present invention are typically applied to an array of Flash EEprom cells in an integrated circuit chip. Figure 12 illustrates schematically an array of individually addressable EEprom cells 1060. Each cell is equivalent to the one shown in Figure 11, having a control gate, source and drain, and an erase gate. The plurality of individual memory cells are organized in rows and columns. Each cell is addressed by selectively energizing its row and column simultaneously. A column 1062, for example, includes a first memory cell 1063, an adjacent second memory cell 1065, and so forth. A second column 1072 includes memory cells 1073, 1075, and so forth. Cells 1063 and 1073 are located in a row 1076, cells 1065 and 1071 in another, adjacent row, and so forth.

Along each row, a word line is connected to all the control gates of the cells in the row. For example, the row 1076 has the word line 1077 and the next row has the word line 1079. A row decoder 1081 selectively connects the control gate voltage V_{CG} on an input line 1083 to all the control gates along a selected word line for a row.

Along each column, all the cells have their sources connected by a source line such as 1091 and all their drains by a drain line such as 1093. Since the cells along a row are connected in series by their sources and drains, the drain of one cell is also the source of the adjacent cell. Thus, the line 1093 is the drain line for the column 1062 as well as the source line for the column 1072. A column decoder 1101 selectively connects the source voltage V_S on an input line 1103 to all the sources and connects the drain voltage V_D on an input line 1105 to all the drains along a selected column.

Each cell is addressed by the row and column in which it is located. For example, if the cell 1075 is addressed for programming or reading, appropriate programming or reading voltages must be supplied to the cell's control gate, source and drain. An address on the internal address bus 1111 is used to decode row decoder 1081 for connecting V_{CG} to the word line 1079 connected to the control gate of the cell 1075. The same address is used to decode column decoder 1101 for connecting V_S to the source line 1093 and V_D to the drain line 1095, which are respectively connected to the source and drain of the cell 1075.

A4 Cont.

One aspect of the present invention, which will be disclosed in more detail in a later section, is the implementation of programming and reading of a plurality of memory cells in parallel. In order to select a plurality of columns simultaneously, the column decoder, in turn, controls the switching of a source multiplexer 1107 and a drain multiplexer 1109. In this way, the selected plurality of columns may have their source lines and drain lines made accessible for connection to V_S and V_D respectively.

Access to the erase gate of each cell is similar to that of the control gate. In one implementation, an erase line such as 1113 or 1115 or 1117 is connected to the erase gate of each cells in a row. An erase decoder 1119 decodes an address on the internal address bus 1111 and selectively connects the erase voltage V_{EG} on input line 1121 to an erase line. This allows each row of cells to be addressed independently, such as the row 1076 being simultaneously (Flash) erased by proper voltages applied to their erase gates through erase line 1113. In this case, the Flash cell consists of one row of memory cells. However, other Flash cell's implementations are possible and most applications will provide for simultaneous erasing of many rows of cells at one time.

Flash EEprom System

The addressable EEprom array 1060 in figure 12 forms part of the larger multi-state Flash EEprom system of the present invention as illustrated in figure 13. In the larger system, an EEprom integrated circuit chip 1130 is controlled by a controller 1140 via an interface 1150. The controller 1140 is itself in communication with a central microprocessor unit 1160.

The EEprom chip 1130 comprises the addressable EEprom array 1060, a serial protocol logic 1170, local power control circuits 1180, and various programming and reading circuits 1190, 1200,1210, 1220, 1230 and 1240.

The controller 1140 controls the functioning of the EEprom chip 1130 by supplying the appropriate voltages, controls and timing. Tables of figures 26 and 27 show typical examples of voltage conditions for the various operational modes of the EEprom cell. The addressable EEprom array 1060 may be directly powered by the controller 1140 or, as shown in figure 13, be further regulated on chip by the local power control 1180. Control and data linkages between the controller 1140 and the chip 1130 are made through the serial in line 1251 and the serial out line 1253. Clock timing is provided by the controller via line 1255.

In a typical operation of the EEprom chip 1130, the controller 1140 will send a serial stream of signals to the chip 1130 via serial in line 1251. The signals, containing control, data, address and timing information, will be sorted out by the serial protocol logic 1170. In appropriate time sequence, the logic 1170 outputs various control signals 1257 to control the various circuits on the chip 1130. It also sends an address via the internal address bus 111 to connect the addressed cell to voltages put out from the controller. In the meantime, if the operation is programming, the data is staged for programming the addressed cell by being sent via a serial data line 1259 to a set of read/program latches and shift registers 1190.

a4 cont'

Read Circuits and Techniques Using Reference Cells

To accurately and reliably determine the memory state of a cell is essential for EEprom operations. This is because all the basic functions such as read, erase verify and program verify depend on it. Improved and novel read circuits 1220 for the EEprom chip 1130 and techniques of the present invention make multi-state EEprom feasible.

As discussed in connection with figure 11, the programmed charge placed on the floating gate 1023' determines the programmed threshold voltage V_{T1} of the cell. Generally, V_{T1} increases or decreases with the amount of negative charge on the floating gate 1023'. The charge can even be reduced to a positive value (depletion mode) where V_{T1} decreases below V_{T2} and even becomes negative. The maximum and minimum values of V_{T1} are governed by the dielectric strength of the device material. The span of V_{T1} defines a threshold voltage window in which memory states may be implemented.

Copending patent application Serial No. 204,175, now patent no. 5,095,344, discloses an EEprom cell with memory states defined within a maximized window of threshold voltage V_{T1}. The full threshold voltage window includes the negative region of the threshold voltage, in addition to the usual positive region. The increased window provides more memory space to implement multi-state in an EEprom cell.

Figures 14 and 15 respectively illustrate the manner in which the threshold voltage window is partitioned for a 2-state memory and a 4-state memory cell. (Of course it is also possible to partition the window for a 3-state memory or even for a continuum of states in an analog, rather than digital memory).

A4 cont.

Referring first to figure 14, the solid curve 1343 shows V_{T1} as a function of programming time. The threshold voltage window is delimited by the minimum and maximum values of V_{T1}, represented approximately by the Erase state level 1345 and the Fully Program state level 1347 respectively. The 2-state memory is implemented by partitioning the window into two halves 1346, 1348 using a breakpoint threshold level 1349. Thus, the cell may be considered to be in memory state 0 (or state 1) if the cell is programmed with a V_{T1} within region 1346 (or region 1348) respectively.

A typical erase/program cycle begins with erase which reduces the threshold voltage of the cell to its Erase state level 1345. Subsequent repetitive programming is used to increase the threshold voltage V_{T1} to the desired level. Rather than continuously applying programming voltages to the addressed cell for some fixed period of time corresponding to the state to which the cell is to be programmed, it is preferable to apply programming voltages in repetitive short pulses with a read operation occurring after each pulse to determine when it has been programmed to the desired threshold voltage level, at which time the programming terminates. The programming voltages and duration of the pulses are such that the pulses advance V_{T1} across the various regions rapidly but each pulse is sufficiently fine to not overshoot any of the regions. This minimizes voltage and field related stresses on the cell, and therefore improves its reliability.

Figure 15A illustrates the 4-state case where the threshold voltage window is partitioned into four regions 1351, 1353, 1355, 1357 by breakpoint levels 1352, 1354, 1356 respectively. The cell is considered to be in state "3" or "2" or "1" or "0" if its V_{T1} is programmed to be within corresponding regions 1351 or 1353 or 1355 or 1357 respectively. A 4-state cell is able to store two bits of data. Thus, the four states may be encoded as (1,1), (1,0), (0,1) and (0,0) respectively.

In general, if each EEprom cell is to store K states, the threshold window must be partitioned into K regions with at least K-1 threshold levels. Thus, only one breakpoint level is required for a 2-state memory cell, and three breakpoint levels are required for a 4-state cell.

In principle, a threshold voltage window may be partitioned to a large number of memory states. For example, for an EEprom device with a maximum threshold window of 16V, it may be partitioned into thirty-two states each within an approximately half volt interval. In practice, prior art EEprom devices have only stored two states or one bit per cell with diminished reliability and life. Apart from operating with a smaller threshold window, prior devices fail to solve two other problems inherent in EEprom devices. Both problems relate to the uncertainty in the amount of charge in the floating gate and hence the uncertainty in the threshold voltage V_{T1} programmed into the cell.

a4 cont.

The first problem has to do with the endurance-related stress the device suffers each time it goes through an erase/program cycle. The endurance of a Flash EEprom device is its ability to withstand a given number of program/erase cycles. The physical phenomenon limiting the endurance of prior art Flash EEprom devices is trapping of electrons in the active dielectric films of the device. During programming, electrons are injected from the substrate to the floating gate through a dielectric interface. Similarly, during erasing, electrons are extracted from the floating gate to the erase gate through a dielectric interface. In both cases, some of the electrons are trapped by the dielectric interface. The trapped electrons oppose the applied electric field in subsequent program/erase cycles thereby causing the programmed V_{T1} to shift to a lower value and the erased V_{T1} to shift to a higher value. This can be seen in a gradual closure in the voltage "window" between the "0" and "1" states of prior art devices as shown in figure 16A. Beyond approximately 1×10^4 program/erase cycles the window closure can become sufficiently severe to cause the reading circuitry to malfunction. If cycling is continued, the device eventually experiences catastrophic failure due to a ruptured dielectric. This typically occurs at between 1×10^6 and 1×10^7 cycles, and is known as the intrinsic breakdown of the device. In prior art EEprom devices the window closure is what limits the practical endurance to approximately 1×10^4 program/erase cycles. This problem is even more critical if multi-state memory is implemented, since more accurate placement of V_{T1} is demanded.

A second problem has to do with the charge retention on the floating gate. The charge on the floating gate tends to diminish somewhat through leakage over a period of time. This causes the threshold voltage V_{T1} to shift also to a lower value over time. Figure 16B illustrates the reduction of V_{T1} as a function of time. Over the life time of the device V_{T1} may shift by as much as 1V. In a multi-state device, this could shift the memory by one or two states.

The present invention overcomes these problems and presents circuits and techniques to reliably program and read the various states even in a multi-state implementation. The memory state of a cell may be determined by measuring the threshold voltage V_{T1} programmed therein. Alternatively, as set forth in co-pending patent application, Serial No. 204,175, now patent no. 5,095,344, the memory state may conveniently be determined by measuring the differing conduction in the source-drain current I_{DS} for the different states. In the 4-state example, figure 15A shows the partition in the threshold voltage window. Figure 15B, on the other hand, illustrates typical values of I_{DS} (solid curves) for the four states as a function of the control gate voltage V_{CG}. With V_{CG} at 5V, the I_{DS} values for each of the four conduction states can be distinguished by sensing with four corresponding current sensing amplifiers in parallel. Associated with each amplifier is a corresponding reference conduction states I_{REF} level (shown as broken curves in figure 16). Just as the breakpoint threshold levels (see figures 14 and 15A) are used to demarcate the different regions in the threshold voltage window, the I_{REF} levels are used to do the same in the corresponding source-drain current window. By comparing with the I_{REF}'s, the conduction state of the memory cell can be determined. Co-pending patent application, Serial No. 204,175, now patent no. 5,095,344, proposes using the same sensing amplifiers and I_{REF}'s for both programming and reading. This provides good tracking between the reference levels (broken curves in figure 15B) and the programmed levels (solid curves in figure 15B).

a4 cont.

In the improved scheme of the present invention, the I_{REF}'s are themselves provided by the source-drain currents of a set of EEprom cells existing on the same chip and set aside solely for this purpose. Thus, they act as master reference cells with their I_{REF}'s used as reference levels for the reading and programming of all other EEprom cells on the same chip. By using the same device as the EEprom cells to act as reference cells, excellent tracking with respect to temperature, voltage and process variations is achieved. Furthermore, the charge retention problem, important in multi-state implementation, is alleviated.

Referring to figure 17A, one such master reference cell 1400 is shown with its program and read paths. The reference cells erase and program module 1410 serves to program or re-program each such reference cell 1400. The module 1410 includes program and erase circuits 1411 with a programming path 1413 connected to the drain of the master reference cell 1400. The circuits 1411 are initiated by addresses decoded from the internal bus 1111 by a program decoder 1415 and an erase decoder 1417 respectively. Accordingly, programming voltages or erasing voltages are selectively supplied each reference cell such as cell 1400. In this way, the reference level in each reference cell may be independently set or reprogrammed. Typically, the threshold level of each reference cell will be factory-programmed to the optimum level appropriate for each batch of chips produced. This could be done by comparison with an external standard reference level. By software control, a user also has the option to reset the reference threshold levels.

A4 cont.

Once the reference threshold voltage V_{T1} or reference drain-source current I_{REF} is programmed into each reference cell 1400, it then serves as a reference for the reading of an addressed memory cell such as cell 1420. The reference cell 1400 is connected to a first leg 1403 of a current sensing amplifier 1410 via a clocked switch 1413. A second leg 1415 of the amplifier is essentially connected to the addressed memory cell 1420 whose programmed conduction state is to be determined. When cell 1420 is to be read, a control signal READ will enable a switch 1421 so that the cell's drain is connected to the second leg 1415. The sense amplifier 1410 supplies voltage via V_{CC} to the drains of both the master reference cell 1400 and the addressed cell 1420. In the preferred embodiment, the amplifier has a current mirror configuration such that any differential in currents through the two legs 1403 and 1415 results in the voltage in the second leg 1415 being pulled up towards V_{CC} or down towards V_S. Thus, the node at the second leg 1415 is respectively HIGH (or LOW) when the source-drain current I_{DS} for the addressed cell 1420 is less (or more) than I_{REF} through the master reference cell 1400. At the appropriate time controlled by a clocked switch 1423, the sensed result at the second leg 1415 may be held by a latch 1425 and made available at an output line 1427. When I_{DS} is less than I_{REF}, a HIGH appears at the output line 1427 and the addressed cell 1420 is regarded as in the same conduction state as the master reference cell 1400.

In the preferred embodiment, a voltage clamp and fast pull-up circuit 1430 is also inserted between the second leg 1415 and the drain 1431 of the addressed cell 1420. The circuit 1430 serves to keep the drain voltage V_D at a maximum of 1.5V - 2.0V when it is charging up in the case of lower I_{DS}. It also prevents V_D from pulling too low in the case of higher I_{DS}.

In general, if each memory cell is to store K states, then at least K-1, or preferably K reference levels need be provided. In one embodiment, the addressed cell is compared to the K reference cells using k sense amplifiers in parallel. This is preferable for the 2-state case because of speed, but may spread the available current too thin for proper sensing in the multi-state case. Thus, for multi-state case, it is preferable to compare the addressed cell with the K reference cells one at a time in sequence.

Figure 17B illustrates more explicitly the multi-state reading configuration. The K reference cells such as 1431, 1433, 1435 are connected to the sense amplifier 1440 via the amplifier's first leg 1441. The connection is time-multiplexed by clocked switches such as 1451, 1453, 1455 respectively. The second leg 1457 of the sense amplifier is connected to the addressed cell as in figure 17A. The sensed signal at the second leg 1457 is time-selectively latched by clocked switches such as 1461, 1463, 1465 onto such latches 1471, 1473, 1475.

Figures 17C(1)-17C(8) illustrate the timing for multi-state read. When the signal READ goes HIGH, a switch 1421 is enabled and the addressed memory cell is connected to the second leg 1457 of the sense amplifier 1440 (figure 17C(1)). The clocks' timing is given in figures 17C(2)-17C(4). Thus, at each clock signal, the sense amplifier sequentially compares the addressed cell with each of the reference cells and latches each results. The latched outputs of the sense amplifier are given in figures 17C(5)-17C(7). After all the K output states of the sense amplifier 1440 are latched, they are encoded by a K-L decoder 1480 ($2^L \geq K$) (figure 17C(8)) into L binary bits.

A4 cont.

Thus, the multiple threshold levels are provided by a set of memory cells which serves as master reference cells. The master reference cells are independently and externally erasable and programmable, either by the device manufacturer or the user. This feature provides maximum flexibility, allowing the breakpoint thresholds to be individually set within the threshold window of the device at any time. By virtue of being the same device as that of the memory cells, the reference cells closely track the same variations due to manufacturing processes, operating conditions and charge retention problems. The independent programmability of each threshold level at will allows optimization and fine-tuning of the partitioning of the threshold window to make multi-state memory viable. Furthermore, it allows post-manufacture configuration for either 2-state or multi-state memory from the same device, depending on user need or device characteristics at the time.

Another important feature of the present invention serves to overcome the problems of endurance-related stress. As explained previously, the erase, program and read characteristics of each memory cell depends on the cumulated stress endured over the number of program/erase cycles the cell has been through. In general, the memory cells are subjected to many more program/erase cycles than the master reference cells. The initially optimized reference levels will eventually become misaligned to cause reading errors. The present underlying inventive concept is to have the reference levels also reflect the same cycling suffered by the memory cells. This is achieved by the implementation of local reference cells in addition to the master reference cells. The local reference cells are subjected to the same program/erase cycling as the memory cells. Every time after an erase operation, the reference levels in the master reference cells are re-copied into the corresponding set of local reference cells. Memory cells are then read with respect to the reference levels of the closely tracking local reference cells. In this way, the deviation in cell characteristics after each program/erase cycle is automatically compensated for. The proper partitioning of the transforming threshold window is therefore maintained so that the memory states can be read correctly even after many cycles.

a4 cont.

Figure 18 illustrates the local cells referencing implementation for Flash EEprom. In the Flash EEprom array 1060 (Fig. 12), each group of memory cells which is collectively erased or programmed is called a sector. The term " Flash sector" is analogous to the term "sector" used in magnetic disk storage devices and they are used interchangeably here. The EEprom array is grouped into Flash sectors such as 1501, 1503 and 1505. While all memory cells in a Flash sector suffer the same cycling, different Flash sectors may undergo different cycling. In order to track each Flash sector properly, a set of memory cells in each Flash sector is set aside for use as local reference cells. For example, after the Flash sector 1503 has been erased, the reference levels in the master reference cells 1507 are re-programmed into the local reference cells associated with the Flash sector 1503. Until the next erase cycle, the read circuits 1513 will continue to read the memory cells within the Flash sector 1503 with respect to the re-programmed reference levels.

Figures 19(1)-19(7) illustrates the algorithm to re-program a sector's reference cells. In particular, figures 19(1)-19(3) relate to erasing the sector's local reference cells to their "erased states". Thus in figure 19(1), a pulse of erasing voltage is applied to all the sector's memory cells including the local reference cells. In figure 19(2), all the local reference cells are then read with respect to the master references cells to verify if they have all been erased to the "erased state". As long as one cell is found to be otherwise, another pulse of erasing voltage will be applied to all the cells. This process is repeated until all the local reference cells in the sector are verified to be in the "erased" state (figure 19(3)).

Figures 19(4)-19(7) relate to programming the local reference cells in the sector. After all the local reference cells in the sector have been verified to be in the "erased" state, a pulse of programming voltage is applied in figure 19(4) only to all the local reference cells. This is followed in figure 19(5) by reading the local reference cells with respect to the master reference cells to verify if every one of the local reference cells is programmed to the same state as the corresponding master reference cell. For those local reference cells not so verified, another pulse of programming voltage is selectively applied to them alone (figure 19(6)). This process is repeated until all the local reference cells are correctly verified (figure 19(7)) to be programmed to the various breakpoint threshold levels in the threshold window.

Once the local reference cells in the sector have been re-programmed, they are used directly or indirectly to erase verify, program verify or read the sector's addressed memory cells.

A4 Cont.

Figure 20A illustrates one embodiment in which the local reference cells are used directly to read or program/erase verify the sector's memory cells. Thus, during those operations, a parallel pair of switches 1525 is enabled by a READ signal and the sense amplifier 1440 will read the sector's addressed memory cells 1523 with respect to each of the sector's local reference cells 1525. During program/erase verify of the local reference cells (as illustrated in figure 19), another parallel pair of switches 1527 enables reading of the local reference cells 1525 relative to the master reference cells 1529.

Figure 20B illustrates the algorithm for using the local reference cells directly to read or program/erase verify the sector's addressed memory cells.

Figure 21A illustrates an alternative embodiment in which the local reference cells are used indirectly to read the addressed memory cells. First the master reference cells are erased and programmed each to one of the desired multiple breakpoint thresholds within the threshold window. Using these master reference thresholds the local reference cells within an erased sector of cells are each programmed to one of the same desired multiple breakpoint thresholds. Next the addressed cells in the sector are programmed (written) with the desired data. The reading sequence for the addressed cells in the sector then involves the steps illustrated in Figure 21A.

First each of the local reference cells 1525 is read relative to the corresponding master reference cell 1531. This is effected by an enabling READ I signal to a switch 1533 connecting the local reference cells 1525 to the second leg 1457 of the sense amplifier 1440 with the master reference 1531 connected to the first leg 1441 of the sense amplifier. Auxiliary current source circuits associated with each master reference cell are now used to optimally bias the current through the first leg 1441 of the sense amplifier to match the current in the second leg 1457. After the bias adjustment operation is completed for all breakpoint threshold levels the addressed cells in the sector are read relative to the bias-adjusted master reference cells. This is effected by disabling READ I to 1533 and enabling READ signal to switch 1535. The advantage of this approach is that any variations in V_{CC}, temperature, cycling fatigue or other effects which may, over time, cause threshold deviations between the master reference cells and the addressed cells is eliminated prior to reading, since the local reference cells (which track threshold deviations of the addressed cells) are used to effectively readjust the breakpoint thresholds of the master reference cells. For example, this scheme permits programming of the addressed cells when the master reference cells are powered with V_{CC}=5.5V and subsequently reading the addressed cells with the master reference cells powered at V_{CC}=4.5V. The difference of 1 volt in V_{CC}, which would normally cause a change in the value of the breakpoint thresholds, is neutralized by using the local reference cells to bias adjust the master reference cells to counteract this change at the time of reading.

A4 cont.

Figures 21B and 21C show in more detail one embodiment of the current biasing circuits such as 1541, 1543, 1545 for the master reference cells 1551, 1553, 1555. Each biasing circuit acts as a current shunt for the current in the master reference cell. For example, the circuit 1541 is tapped to the drain of the master reference cell 1551 through the line 1561. It modifies the current in line 1562 to the sense amplifier (first leg) either by sourcing current from V_{CC} or draining current to V_{SS}. In the former case, the current in the line 1562 is reduced, and otherwise for the latter case. As biasing is being established for the master reference 1551, any inequality in the currents in the two legs of the sense amplifier can be communicated to outside the chip. This is detected by the controller (see figure 13) which in turn programs the biasing circuit 1541 via the internal address bus 1111 to subtract or add current in the line 1562 in order to equalize that of the local reference.

Figure 21C illustrates an embodiment of the biasing circuit such as the circuit 1541. A bank of parallel transistors such as 1571, 1573, 1575 are all connected with their drains to V_{CC}, and their sources via switches such as 1581, 1583, 1585 to the line 1561. By selectively enabling the switches, different number of transistors may be used to subtract various amount of current from line 1562. Similarly, another bank of parallel transistors such as 1591, 1593, 1595 are all connected with their sources to V_{SS}, and their drains via switches such as 1601, 1603, 1605 to the line 1561. By selectively enabling the switches, a different number of transistors may be used to add a various amount of current to line 1562. A decoder 1609 is used to decode address from the internal address bus 1111 to selectively enable the switches. The enabling signals are stored in latches 1611, 1613. In this way every time a sector is read, the master reference cells are re-biased relative to the local reference cells, and used for reading the memory cells in the sector.

Figures 21D(1)-21D(4) illustrate the read algorithm for the alternative embodiment. The sector must previously have had its local reference cells programmed and verified relative to the master reference cells (figure 21D(1)). Accordingly, each of the master reference cells is then read relative to the local reference cells (figure 21D(2)). The master reference cells are biased to equalize the current to that of the corresponding local reference cells (figure 21D(3)). Subsequently, the memory cells in the sector are read relative to the biased master reference cells(figure 21D(4)).

The read circuits and operation described are also employed in the programming and erasing of the memory cells, particularly in the verifying part of the operation. As described previously, programming is performed in small steps, with reading of the state programmed in between to verify if the desired state has been reached. As soon as the programmed state is verified correctly, programming stops. Similarly, erasing is performed in small steps, with reading of the state of erase in between to verify if the "erased" state has been reach. Once the "erased" state is verified correctly, erasing stops.

A4 cont'd

As described previously, only K-1 breakpoint threshold levels are required to partition the threshold window into K regions, thereby allowing the memory cell to store K states. According to one aspect of the present invention, however, in the multi-state case where the threshold window is more finely partitioned, it is preferable to use K threshold levels for K state. The extra threshold level is used to distinguish the "erased" state from the state with the lowest threshold level. This prevents over-erasing and thus over-stressing the cell since erasing will stop once the "erased" state is reached. The selective inhibition of individual cells for erase does not apply to the Flash EEprom case where at least a sector must be erased each time. It is suitable for those EEprom arrays where the memory cells can be individually addressed for erase.

According to another feature of the invention, after a memory cell has been erased to the "erased" state, it is programmed slightly to bring the cell to the state with the lowest threshold level (ground state) adjacent the "erased" state. This has two advantages. First, the threshold levels of the ground state of all the memory cells, being confined between the same two breakpoint threshold levels, are well-defined and not widely scattered. This provide an uniform starting point for subsequent programming of the cells. Secondly, all cells get some programming, thereby preventing those cells which tend to have the ground state stored in them, for example, from losing track with the rest with regard to program/erase cycling and endurance history.

On Chip Program Verify

As mentioned before, programming of an EEprom cell to a desired state is preferably performed in small steps starting from the "erase" state. After each programming step, the cell under programming is read to verify if the desired state has been reached. If it has not, further programming and verifying will be repeated until it is so verified.

Referring to the system diagram illustrated in figure 13, the EEprom chip 1130 is under the control of the controller 1140. They are linked serially by the serial in line 1251 and serial out line 1253. In prior art EEprom devices, after each programming step, the state attained in the cell under programming is read and sent back to the controller 1140 or the CPU 1160 for verification with the desired state. This scheme places a heavy penalty on speed especially in view of the serial link.

A4 cont'

In the present invention, the program verification is optimized by programming a chunk (typically several bytes) of cells in parallel followed by verifying in parallel and on chip. The parallel programming is implemented by a selective programming circuit which disables programming of those cells in the chunk whose states have already been verified correctly. This feature is essential in a multi-state implementation, because some cells will reach their desired state earlier than others, and will continue past the desired state if not stopped. After the whole chunk of cells have been verified correctly, logic on chip communicates this fact to the controller, whereby programming of the next chunk of cells may commence. In this way, in between each programming step data does not need to be shuttled between the EEprom chip and the controller, and program verification speed is greatly enhanced.

Figure 22 illustrates the program and verify paths for a chunk of n cells in parallel. The same numerals are used for corresponding modules in the system diagram of figure 13. The EEprom array 1060 is addressed by N cells at a time. For example, N may be 64 cells wide. In a 512 bytes Flash sector, consisting of 4 rows of 1024 cells, there will be 64 chunks of 64 cells. The source multiplexer 1107 selectively connects the N sources of one addressed chunk of cells to the source voltage V_S in line 1103. Similarly, the drain multiplexer 1109 selectively makes the N drains of the chunk accessible through an N-channel data path 1105. The data path 1105 is accessed by the program circuit with inhibit 1210 during programming and by read circuits 1220 during reading, program verifying or erase verifying.

Referring again to the system diagram in figure 13, programming is under the control of the controller 1140. The data to be programmed into the sector is sent chunk by chunk. The controller first sends a first chunk of N*L serial data bits together with addresses, control and timing information to the EEprom chip 1130. L is the number of binary bits encoded per memory cell. For example, L=1 for a 2-state cell, and L=2 for a 4-state cell. Thus if N=64 and L=2, the chunk of data bits will be 128 bits wide. The N*L data bits are stored in latches and shift registers 1190 where the serial bits are converted to N*L parallel bits. These data will be required for program verify in conjunction with the read circuits 1220, bit decoder 1230, compare circuit 1200 and the program circuit with inhibit 1210.

A4 cont'd

The program algorithm for a chunk of N cells is best described by referring to both the system diagram of figure 13 and figures 23(1)-23(7) which illustrate the algorithm itself. As mentioned in an earlier section, prior to programming the sector, the whole sector must be erased and all cells in it verified to be in the "erased" state (figure 23(1)). This is followed in figure 23(2) by programming the sector local reference cells (as shown in figures 19(1)-(3)). In figure 23(3), the N*L bits of parallel data is latched in latches 1190. In figure 23(4), the read circuits 1220 access the N-channel data path 1105 to read the states in the N chunk of cells. The read algorithm has already been described in conjunction with figure 20B or figure 21D. The N-cell reads generates N*K (K=number of states per cell) output states. These are decoded by bit decoder 1230 into N*L binary bits. In figure 23(5), the N*L read bits are compared bit by bit with the N*L program data bits from latches 1190 by compare circuit 1200. In figure 23(6), if any read bit fails to compare with the program data bit, a further programming voltage pulse from the program circuit 1210 is applied simultaneously to the chunk of cells. However, an inhibit circuit within the program circuit 1210 selectively blocks programming to those cells whose bits are correctly verified with the programmed data bits. Thus, only the unverified cells are programmed each time. Programming and verification are repeated until all the cells are correctly verified in figure 23(7).

Figure 24 shows one embodiment of the compare circuit 1200 of figure 13 in more detail. The circuit 1200 comprises N cell compare modules such as 1701, 1703, one for each of the N cells in the chunk. In each cell compare module such as the module 1701, the L read bits (L=number of binary bits encoded for each cell) are compared bit by bit with the corresponding program data bits. This is performed by L XOR gates such as 1711, 1713, 1715. The output of these XOR gates pass through an NOR gate 1717 such that a "1" appears at the output of NOR gate 1717 whenever all the L bits are verified, and a "0" appears when otherwise. When the control signal VERIFY is true, this result is latched to a latch 1721 such that the same result at the output of NOR gate 1717 is available at the cell compare module's output 1725. The compare circuit 1200 performs the comparisons of L bits in parallel. The N compare module's outputs such as 1725, 1727 are available at an N-channel output line 1731 to be fed to the program circuit with inhibit 1210 of figure 13.

A4 cont.

At the same time, the N outputs such as 1725, 1727 are passed through an AND gate 1733 so that its single output 1735 results in a "1" when all N cells are verified and a "0" when otherwise. Referring also to figure 13, the single output 1735 is used to signal the controller 1140 that all N cells in the chunk of data have been correctly verified. The signal in output 1735 is sent through the serial out line 1253 via AND gate 1240 during a VERIFY operation.

At power-up or at the end of program/verify of a chunk of data, all cell compare module's outputs such as 1725, 1727 are reset to the "not-verified" state of "0". This is achieved by pulling the node 1726 to V_{SS} (0V) by means of the RESET signal in line 1727 to a transistor 1729.

Figure 25 shows one embodiment of the program circuit with inhibit 1210 of figure 13 in more detail. The program circuit 1210 comprises N program with inhibit modules such as 1801, 1803. As illustrated in the tables of figures 26 and 27, in order to program the N cells, a voltage V_{PD} must be applied to each of the N cells' drain and a voltage V_{PG} applied to the control gates. Each program module such as 1801 serves to selectively pass V_{PD} on a line 1805 to one of the drains through the one of the N-channel data path 1105. Since V_{PD} is typically about 8V to 9V which is higher than V_{CC}, the latter cannot be used to turn on the transistor switch 1807. Rather the higher voltage V_{CG} (about 12V) is used to enable switch 1807. V_{CG} in line 1801 is itself enabled by an AND gate when both the program control signal PGM in line 1813 is true and the signal in line 1731 is a "0". Since the signal in line 1731 is from the output of the cell compare module 1701 shown in figure 24, it follows that V_{PD} will be selectively passed onto those cells which are not yet verified. In this way, every time a programming pulse is applied, it is only applied to those cells which have not yet reached their intended states. This selective programming feature is especially necessary in implementing parallel programming and on chip verification in the multi-state case.

Variable Control of Voltage to the Control Gate

The system diagram of figure 13 in conjunction with figures 26 and 27 illustrate how various voltages are applied to the EEprom array 1060 to perform the basic functions of the EEprom.